ExTaSem! Extending, Taxonomizing and Semantifying Domain Terminologies

Luis Espinosa-Anke*, Horacio Saggion*, Francesco Ronzano* and **Roberto Navigli** †
* DTIC - TALN Research Group, Universitat Pompeu Fabra, Carrer Tànger 122-134, 08018 Barcelona (Spain)
† Department of Computer Science, Sapienza University of Rome, Viale Regina Elena, 295, Rome (Italy)
*{luis.espinosa, horacio.saggion, francesco.ronzano}@upf.edu, †navigli@di.uniroma1.it

Abstract

We introduce EXTASEM!, a novel approach for the automatic learning of lexical taxonomies from domain terminologies. First, we exploit a very large semantic network to collect thousands of in-domain textual definitions. Second, we extract (hyponym, hypernym) pairs from each definition with a CRF-based algorithm trained on manually-validated data. Finally, we introduce a graph induction procedure which constructs a full-fledged taxonomy where each edge is weighted according to its domain pertinence. EXTASEM! achieves state-of-the-art results in the following taxonomy evaluation experiments: (1) Hypernym discovery, (2) Reconstructing gold standard taxonomies, and (3) Taxonomy quality according to structural measures. We release weighted taxonomies for six domains for the use and scrutiny of the community.

Introduction

Question Answering and Reasoning, as well as other applications in Artificial Intelligence and Natural Language Processing (NLP), can benefit dramatically from semantic knowledge. Many approaches for creating and formalizing this knowledge are based on domain ontologies, whose backbone are *lexical taxonomies* (Navigli, Velardi, and Faralli 2011). The term taxonomy is used to refer to graph-like hierarchical structures where concepts are nodes organized over a predefined merging or splitting criterion (Hwang, Grauman, and Sha 2012). For example, WordNet (Miller et al. 1990) groups words into sets of super and subordinate (is-a) relations. Taxonomies have proven beneficial for tasks like Question Answering (Harabagiu, Maiorano, and Pasca 2003) or textual entailment (Glickman, Dagan, and Koppel 2005).

Prominent examples of projects leveraging taxonomic organization of knowledge are lexical databases like WordNet, or knowledge-oriented endeavours aimed at taxonomizing large information repositories, such as YAGO (Suchanek, Kasneci, and Weikum 2007), WikiTaxonomy (Ponzetto and Strube 2008) or the Wikipedia Bitaxonomy (WIBI) (Flati et al. 2014). More recently, the first Semeval Task on Taxonomy Learning Evaluation (Bordea et al. 2015) aimed at providing a common evaluation benchmark for taxonomy learning.

Previous methods for inducing taxonomic relations can be (broadly) classified into linguistic or statistic. Linguistic methods are those that, extending Hearst's patterns (Hearst 1992), exploit linguistic evidence for unveiling hypernym relations (Kozareva and Hovy 2010; Navigli, Velardi, and Faralli 2011; Flati et al. 2014; Luu Anh, Kim, and Ng 2014). Other approaches are based purely on statistical evidence and graph-based measures (Fountain and Lapata 2012; Alfarone and Davis 2015). However, none of these approaches addressed explicitly the problem of ambiguity and semantically-motivated domain pertinence, albeit a few cases in which all this was tackled tangentially (Kozareva and Hovy 2010; Velardi, Faralli, and Navigli 2013).

EXTASEM! is designed to bridge the gap between relation extraction and graph construction, on one hand, and domain pertinence on the other. Starting from a list of domain terms, EXTASEM! induces a full-fledged taxonomy by leveraging a large semantic network, from which *high quality knowledge* in the form of textual definitions is retrieved for each domain. Then, (hyponym, hypernym) pairs are extracted via a Conditional Random Fields (CRF) based sequential classifier. Finally, a state-of-the-art vector space representation of individual word senses is exploited for constructing a domain taxonomy only made up of semantically pertinent edges[1]. Our approach does not require a step for graph pruning or trimming, a must in some of the systems mentioned above.

In terms of taxonomy evaluation, EXTASEM! is able to reliably reconstruct gold standard taxonomies of interdisciplinary domains such as Science, Terrorism or Artificial Intelligence, as well as more specific ones like Food or Equipment. In addition, it has the capacity to extend and semantify an input taxonomy, i.e. increase its size and link many of its nodes to a reference sense inventory.

Related Work

Building up on the pioneering work by (Hearst 1992) for hypernym discovery, later methods have leveraged linguistic regularities as a first step for taxonomy learning. Some of these works include KnowItAll (Etzioni et al. 2005),

[1] Taxonomies available at http://taln.upf.edu/extasem.

designed over tamplates for harvesting candidate instances which are afterwards ranked via Mutual Information. Another well-known contribution exploits syntactic evidence together with a probabilistic framework (Snow, Jurafsky, and Ng 2006), using WordNet hypernym relations to learn syntactic dependencies and introduce them as features into a logistic regression classifier. Taxonomies can also be constructed combining syntactic dependencies and structural information present in Wikipedia such as hyperlinks (Flati et al. 2014). Finally, (Kozareva and Hovy 2010) introduce the *double anchored* method, which retrieves a priori disambiguated (hyponym, hypernym) pairs, as a first stage of a graph-based taxonomy learning algorithm.

Taxonomy learning can also be cast as a clustering problem. For instance, (Hjelm and Buitelaar 2008) observe that multilingual distributional evidence can be effectively used for clustering terms hierarchically using the k-means algorithm. Furthermore, (Liu et al. 2012) propose a "knowledge + context" hierarchical clustering approach, where key domain terms are extracted from a general-purpose Knowledge Base (KB) and afterwards the web is used as source for contextual evidence. Contextual evidence is also used in (Luu Anh, Kim, and Ng 2014), who assign a taxonomic relation to concept pairs according to predefined syntactic relations over dependency trees, e.g. if two terms appear in a *Subject-Verb-Object* pattern.

As for graph-based systems, (Fountain and Lapata 2012) leverage hierarchical random graphs, modelling the probability of a given edge to exist in a target taxonomy. In the case of *OntoLearn ReLoaded* (Velardi, Faralli, and Navigli 2013), syntactic evidence, distributional information and domain filtering are combined with a graph-based algorithm to achieve an optimal branching for an initially dense taxonomy. Finally, (Alfarone and Davis 2015) combine inference based on distributional semantics with an approach to remove incorrect edges, again aiming at an optimal non-reduntant branching.

Resources

EXTASEM! operates on the back of two semantic knowledge sources: BABELNET (Navigli and Ponzetto 2010) and SENSEMBED (Iacobacci, Pilehvar, and Navigli 2015).

BabelNet We leverage BABELNET[2] mainly as a definition source. Our choice stems from the fact that it currently constitutes the largest single multilingual repository of named entities and concepts, containing around 14M synsets enriched with a set of *definitions*, available thanks to the seamless integration of resources such as Wikipedia, OmegaWiki, Wiktionary, Wikidata and WordNet.

SENSEMBED EXTASEM! takes advantage of a vector space representation of items, which is exploited to compute the domain pertinence of a candidate edge to the taxonomy. Current representations like word embeddings (Mikolov, Yih, and Zweig 2013) or multi sense models (Neelakantan et al. 2014) associate one or more vectors to individual words. However, these are not represented in any reference sense inventory, which we observe can provide significant support in term of is-a relations and domain pertinence. Hence, since semantic information is crucial in the taxonomy learning task, we leverage SENSEMBED, a knowledge-based approach for obtaining latent continuous representations of individual word senses. It exploits the structured knowledge of a large sense inventory along with the distributional information gathered from text corpora. SENSEMBED vectors are trained on the English Wikipedia, with BABELNET as a reference sense inventory.

[2]http://babelnet.org

Method

In this section, we describe the pipeline of EXTASEM! and the resources enabling its semantic properties. Let $\mathbf{I}_\varphi$ be a set of terms in domain φ, where:
$\varphi \in \{\text{Food}, \text{Equipment}, \text{Science}, \text{Chemical}, \text{AI}, \text{Terrorism}\}$[3] and let $\mathbf{T}_\varphi$ be the final domain taxonomy, which can be described as a directed acyclic graph. The root node of the taxonomy corresponds with a generic umbrella term of the target domain φ. This paper describes the procedure to learn $\mathbf{T}_\varphi$ from $\mathbf{I}_\varphi$.

Domain Definition Harvesting

Following previous work in Definition Extraction (Saggion 2004; Navigli and Velardi 2010), EXTASEM! extracts candidate hypernyms of terms by mining textual definitions retrieved from reliable knowledge sources. In this way we can focus on the semantic coherence and the completeness of the taxonomy we build with respect to both the addition of novel terms and edges and the evaluation of their quality against reference sense inventories. Moreover, by gathering definitions from reliable knowledge sources we reduce the risk of semantic drift in our taxonomy and the need of costly and often imprecise pruning approaches. These approaches are usually adopted when evidence is harvested from non-curated data like the web (Kozareva and Hovy 2010) or the output of Open Information Extraction (OIE) systems (Alfarone and Davis 2015).

The first component of the EXTASEM! pipeline is the Domain Definition Harvesting (DDH) module. Given a domain terminology $\mathbf{I}_\varphi$, the DDH module collects a corpus of domain definition sentences D_φ retrieved from BABELNET that constitutes our *global definition repository*.

The DDH module consists of two sequential phases (see Figs. 1 and 2): the *Computation of the Domain Pertinence Score of Wiki-Categories* (DDH-CatDPScore) and the *Domain Definitions Gathering* (DDH-DefGath). The DDH-CatDPScore generates a list of Wikipedia Categories, each one characterized by a score that quantifies its pertinence to the domain of the input terminology $\mathbf{I}_\varphi$. Then, the DDH-DefGath prunes further this list of Wikipedia Categories with respect to their domain relevance and semantic coherence and, then, exploits the pruned Category list to populate the corpus of domain definition sentences (D_φ). Hereafter we describe each phase in detail.

[3]See the Evaluation section for the motivation behind the choice of these domains.

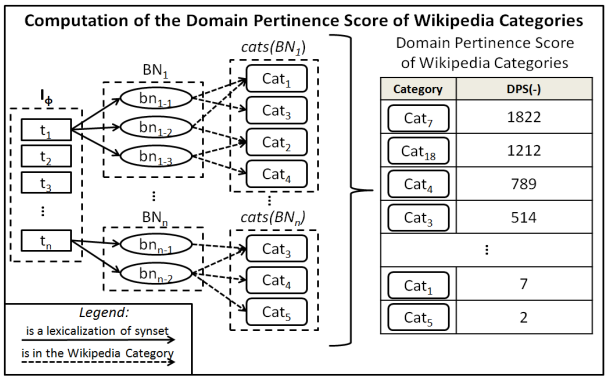

Figure 1: DDH: DPS computation phase.

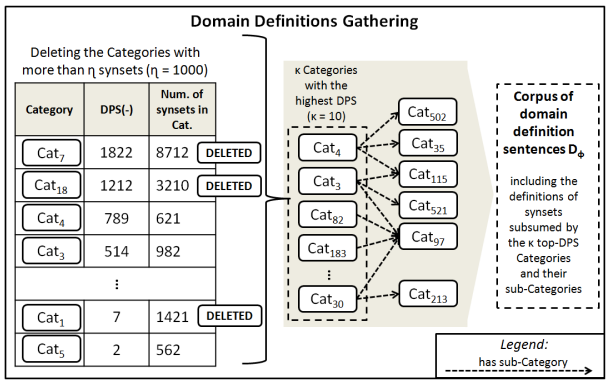

Figure 2: DDH: Domain Definitions Gathering phase.

DDH-CatDPScore (see Fig. 1): for each term τ belonging to the input domain terminology $\mathbf{I}_\varphi$, we collect the BABELNET synsets BN_τ that include the term τ as one of their lexicalizations. Then, exploiting the Wikipedia Bitaxonomy (Flati et al. 2014) integrated in BABELNET, for each set of BABELNET synsets BN_τ, we compute $cats(BN_\tau)$, i.e. the set of Wikipedia Categories that include at least one BABELNET synset in BN_τ. We compute the Domain Pertinence Score (DPS) of each Wikipedia Category CAT_n:

$$DPS(\text{CAT}_n) = \sum_{\tau \in \mathbf{I}_\varphi} \begin{cases} 1 & \text{if } \text{CAT}_n \in cats(BN_\tau) \\ 0 & \text{if } \text{CAT}_n \notin cats(BN_\tau) \end{cases}$$

The DPS of each Category is equal to the number of terms τ that represent one of the lexicalizations of a BABELNET synset included in the same Category (thus belonging to BN_τ). We rely on the intuition that the greater the DPS of a Category is, the higher is the relevance of that Category to the domain of the terminology $\mathbf{I}_\varphi$. The output of the DDH-CatDPScore phase is the list of all Wikipedia Categories that have a DPS greater than zero.

DDH-DefGath (see Fig. 2): from the list of Wikipedia Categories that have a DPS greater than zero, we filter out those that include more than η synsets. We applied this procedure since we noted that often Wikipedia Categories that include large amounts of synsets are one-size-fits-all repositories. These Categories may not be relevant to characterize our domain of interest since they often group huge amounts of semantically heterogeneous synsets, thus showing low semantic coherence. Examples of these Categories are: *Living People* or *English Language Films*. As a consequence of the analysis of several cases, we empirically set the Category exclusion threshold η to 1000. From the filtered list of Categories, we select the κ Categories with the highest DPS (top-κ). From each BABELNET synset that is included in a top-κ Category or one of its sub-Categories, we collect all the full-sentence definitions in BABELNET (which includes Wikipedia, WikiData, OmegaWiki, WordNet, and Wiktionary definitions). In all the experiments reported in this paper, we set κ equal to 10. The set of full-sentence definitions we collect constitutes our corpus of domain definition sentences D_φ.

Hypernym Extraction

A core component of our pipeline is the Hypernym Extraction (HE) module. Given a textual definition $d_\tau \in D_\varphi$, we obtain the *longest and most specific* hypernym of τ. Then, exploiting the syntactic structure of each multiword hypernym, we propose a hypernym-decomposition step for increasing the depth of the graph, which is preferred in taxonomy learning (Navigli, Velardi, and Faralli 2011).

We cast HE as a sequential classification problem where, for each word, our model predicts whether such word is at the beginning of (B), inside of (I) or outside (O) a hypernym (single or multiword). Recall that at this stage we have extracted many definitions in the form of both full definitional sentences and glosses. For full sentence definitions (like the ones in Wikipedia), we trained a model with the WCL corpus, a manually validated dataset (Navigli, Velardi, and Ruiz-Martínez 2010). It contains about 2000 sentences, including definitions annotated following the *genus et differentia* model, and what the authors called *syntactically plausible false definitions*, i.e. sentences that include a domain term but are not properly defining it. For dictionary-style glosses, we manually annotated the hypernym of 500 glosses from WordNet. These glosses came from domains unrelated to the ones in which the experiments were performed.

Prior to training, data is preprocessed and parsed with a dependency parser (Bohnet 2010). We follow (Espinosa-Anke, Saggion, and Ronzano 2015) and apply a combination of linguistic and statistic features. They are used to train a Conditional Random Fields (Lafferty, McCallum, and Pereira 2001) model[4] with a word-level context window of [3, -3]. This window is designed to capture *definition triggers* (e.g. *is a* or *constitutes a*) as well as recurrent key phrases at definiens position (e.g. *which is considered* or *famous for*).

The model is applied to D_φ to extract a set $\mathbf{H}_\varphi$ of (hyponym, hypernym) pairs. At this stage, τ may be associated with more than one hypernym, as we may extract several candidates from different definition sources. For ex-

[4]CRF++: http://crfpp.googlecode.com/

ample, for τ = TRUFFLE, extracted candidates are CONFECTION, GANACHE CENTER, and CHOCOLATE CANDY. Note that GANACHE CENTER is a wrong hypernym for TRUFFLE, and will eventually be pruned out.

Fine-Graining Hyponym - Hypernym Pairs

We propose a *hypernym decomposition* heuristic over the syntactic dependencies in a definition d_τ. In dependency grammar, a sentence is represented as a lexical tree where words depend on each other in various ways (Francom and Hulden 2008). We exploit this linguistic structure to: (1) Extract from the sentence the dependency subtree rooted at the head of the hypernym candidate; (2) Remove one modifier at a time until the hypernym candidate consists only of one token. A syntactic constraint is introduced to retain only relevant modifiers, i.e. only nouns, adjectives and verbs are kept. This procedure outputs a finer-grained set of relations, denoted as $\mathbf{H}'_\varphi$. For example, JAPANESE SNACK FOOD $\mapsto$ {JAPANESE SNACK FOOD, SNACK FOOD, FOOD}.[5]

After the hypernym decomposition step, we construct a set of *candidate paths* P^φ from $\mathbf{H}'_\varphi$. A candidate path $p^\varphi_\tau \in P^\varphi$ is defined as a path from a term node τ to the root node φ, and includes as intermediate nodes those created during the syntactic decomposition step. From our previous example, {JAPANESE SNACK FOOD, SNACK FOOD, FOOD} $\mapsto$ {JAPANESE SNACK FOOD $\rightarrow$ SNACK FOOD $\rightarrow$ FOOD}[6]. In the following section, we explain how EXTASEM! constructs a domain-pertinent taxonomy from P^φ.

Path Weighting and Taxonomy Induction

We expect *good paths* to be relevant to the domain. We could model this relevance in terms of syntactic evidence (Luu Anh, Kim, and Ng 2014), frequency in knowledge generated by OIE systems (Alfarone and Davis 2015) or the web (Kozareva and Hovy 2010). However, recent work in vectorial representations of semantically-enhanced items has shown state-of-the-art performance in several word similarity and word relatedness tasks (Camacho-Collados, Pilehvar, and Navigli 2015). This suggests that these representations may be much more suitable for our semantics-intensive path weighting policy. Thus, we incorporate a module based on SENSEMBED, which operates on the back of a sense inventory S with a corresponding vector space Γ.

We model the relevance of p^φ_τ to φ (e.g. Food or Chemical) by computing its *domain pertinence*. This is given by the weighting function $w(\cdot)$, computed as the cumulative semantic similarity between each node $n \in p^\varphi_\tau$ and φ. We first gather all the available senses in S of both n and φ, namely $S(n) = \{s_{n^1}, ..., s_{n^m}\}$ and $S(\varphi) = \{s_{\varphi^1}, ..., s_{\varphi^z}\}$, and we retrieve from Γ the corresponding sets of vectors $V(n) = \{v_{n^1}, ..., v_{n^m}\}$ and $V(\varphi) = \{v_{\varphi^1}, ..., v_{\varphi^z}\}$. Our aim now is to assign to n the closest sense to φ so that, for instance, for the node *apple*, the correct sense in the Food domain will be that of the fruit, and not that of the company.

[5] A manual analysis over a random sample of 100 edges in the AI domain showed that compositionality failed in less than 6% of the cases.

[6] Henceforth, we denote edges as term$\rightarrow$hypernym.

Next, we compare each possible pair of senses and select the one maximizing the *cosine similarity* COS between their corresponding vectors:

$$\cos(n, \varphi) = \max_{v_n \in V(n), v_\varphi \in V(\varphi)} \frac{v_n \cdot v_\varphi}{||v_n|| ||v_\varphi||}$$

Then, we weigh each path as follows:

$$w(p^\varphi_\tau) = \sum_{l \in L(p^\varphi_\tau)} \cos(l, \varphi)$$

where $L(p^\varphi_\tau)$ is the set of *linkable nodes* in a path, i.e. those nodes with at least one vector representation associated with them.

This yields P^φ_W, a weighted set of candidate edges. For instance, {(MIKADO$\rightarrow$JAPANESE SNACK FOOD), (JAPANESE SNACK FOOD$\rightarrow$SNACK FOOD), (SNACK FOOD$\rightarrow$FOOD)}$_{w=0.3}$. Finally, the taxonomy induction module generates a full-fledged semantified taxonomy $\mathbf{T}_\varphi$ with many intermediate nodes which were not present in $\mathbf{I}_\varphi$, as well as a large number of novel non-redundant edges. This last step is described in Alg. 1. We empirically set a threshold θ to .135, and apply it over all domains.

Algorithm 1 Taxonomy Induction

Input: Threshold θ, weighted paths P^φ_W
Output: Disambiguated domain taxonomy $\mathbf{T}_\varphi$

```
/*A(term, T_φ) denotes the set of ancestors of term in T_φ */
T_φ = ∅
for ρ^φ_τ ∈ P^φ_W do
    if w(ρ^φ_τ) > θ then
        for (term, hyp) ∈ ρ^φ_τ do
            if hyp ∉ A(term, T_φ) then
                T_φ = T_φ ∪ {term → hyp}
return T_φ
```

Evaluation

Evaluating the quality of lexical taxonomies is an extremely difficult task, even for humans (Kozareva, Hovy, and Riloff 2009). This is mainly because there is not a single way to model a domain of interest (Velardi, Faralli, and Navigli 2013), and even a comparison against a gold standard may not reflect the true quality of a taxonomy, as gold standard taxonomies are not complete. This is especially relevant in multidisciplinary and evolving domains such as Science (Bordea et al. 2015). Thus, we evaluated EXTASEM! from two different standpoints, namely: (1) Reconstructing a gold-standard taxonomy; and (2) Taxonomy quality. We used the following data for our experiments:

1. **TExEval 2015**: We evaluated on Semeval-2015 Task 17 (TExEval) domains: Science (*sci.*), Food (*food*), Equipment (*equip.*) and Chemical (*chem.*). For each domain, two terminologies and their corresponding gold standard taxonomies were available. Such gold standards came from both domain-specific sources (e.g. for *chem.*, the

ChEBI taxonomy[7]) and the WordNet subgraph rooted at the domain concept (e.g. the WordNet subtree rooted at *chemical* in the case of *chem.*). Note that since WordNet is integrated in BABELNET, evaluation over WordNet gold standard would artificially favour our approach, so we decided to only evaluate on the domain-specific taxonomies. We compared our results against the taxonomies produced by task participants.

2. **Additional *multidisciplinary* domains**: We assessed the EXTASEM! taxonomies in the domains of Artificial Intelligence (*AI*) (Velardi, Faralli, and Navigli 2013) and Terrorism (*terr.*) (Luu Anh, Kim, and Ng 2014). For the same fairness reason as above, we avoid domains covered in previous work where the gold standard comes from WordNet, such as Animals, Plants and Vehicles, used in (Velardi, Faralli, and Navigli 2013; Kozareva and Hovy 2010; Alfarone and Davis 2015).

	Food			Science			Chem.			Equip.		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
INRIASAC	.18	.51	.27	.17	.44	.25	.08	.09	**.09**	.26	.49	.34
LT3	.28	.29	.29	.40	.38	**.39**	-	-	-	.70	.32	.44
ntnu	.07	.05	.06	.05	.04	.04	.02	.002	.001	.01	.006	.009
QASSIT	.06	.06	.06	.20	.22	.21	-	-	-	.24	.24	.24
TALN-UPF	.03	.03	.03	.07	.25	.11	-	-	-	.14	.15	.15
USAARWLV	.15	.26	.20	.18	.37	.24	.07	.09	.08	.41	.36	.39
EXTASEM!	.28	.66	**.39**	.27	.32	.29	.05	.02	.03	.51	.56	**.54**

Table 1: Comparative edge-level Precision, Recall and F-measure scores. Refer to (Bordea et al. 2015) for a description of each of the systems listed.

Reconstructing a Gold Standard

Experiment 1 - TExEval 2015 The taxonomies generated by EXTASEM! are compared against participant systems in TExEval. The evaluation criterion in this experiment is to assess how well systems can replicate a gold standard in any of the four evaluated domains. This is done via Precision, Recall and F-Score at edge level.

EXTASEM! ranks first in half of the domains (Table 1), and second and third in Science and Chemical respectively. Note that if we average the results of all the systems participating in this experiment across the four domains, our approach ranks first (F=0.31, the second best system being LT3 with F=0.28).

Experiment 2 - Evaluation of a Subsample The Cumulative Fowlkes&Mallows Measure (CFM) (Velardi, Faralli, and Navigli 2013) has become a de-facto standard for evaluating lexical taxonomies against ground truth. It was introduced as a rework of the original Fowlkes&Mallows measure (Fowlkes and Mallows 1983), and was used as one of the evaluation criteria in TExEval 2015. This measure assigns a score between 0 and 1 according to how well a system clusters similar nodes at different cut levels.

[7]https://www.ebi.ac.uk/chebi/

In this experiment, we took advantage of extensive human input, and asked domain experts to reconstruct a sample of 100 concepts from taxonomies produced by EXTASEM!. The reason for having a sample of 100 terms is that it is a compact enough sample to avoid the "messy organization" previous authors have reported (Velardi, Faralli, and Navigli 2013; Kozareva and Hovy 2010), while being a larger sample than experiments performed similarly, e.g. in (Fountain and Lapata 2012), where the terminologies given to human judges were only of 12 terms.

For each 100-term sample, a domain expert was asked to order hierarchically as many concepts as possible, but was allowed to leave out any node if it was considered noisy. We used these expert taxonomies as gold standard. We also evaluated a baseline method based on substring inclusion consisting in creating a hyponym→hypernym pair between two terms if one is prefix or suffix substring of the other. Table 2 shows results in terms of edge overlap (RECALL) and CFM. The agreement between EXTASEM! and human experts was high, performing much better than the baseline.

	Baseline		**EXTASEM!**	
	RECALL	**CFM**	**RECALL**	**CFM**
Food	0.49	0.02	0.79	0.50
Science	0.22	0.01	0.57	0.64
Equip.	0.43	0.01	0.773	0.506
Terr.	0.54	0.07	0.697	0.274
AI	0.51	0.02	0.771	0.497

Table 2: CFM for domain 100-term gold standard comparison.

Taxonomy Quality

Experiment 1 - Structural Evaluation

According to (Bordea et al. 2015), the purpose of taxonomy structural evaluation is to: (1) Quantify its size in terms of nodes and edges; (2) Assess whether all components are connected; and (3) Quantify semantic richness in terms of proportion of intermediate nodes versus leaf nodes (which are considered less important). Thus, we compare automatic taxonomies produced by EXTASEM! with gold standard taxonomies from TExEval 2015 (TEXE) in all domains, as well as automatic taxonomies produced in Artificial Intelligence (*AI*) (Velardi, Faralli, and Navigli 2013) and Terrorism (*terr.*) (Luu Anh, Kim, and Ng 2014). We evaluated over these parameters: Number of nodes (NODES); number of edges (EDGES); number of connected components (C.C); number of intermediate nodes, i.e. those which are neither root or leaf nodes (I.N); maximum depth of the taxonomy (MD); and average depth (AD).

EXTASEM! produces bigger taxonomies with more intermediate nodes in three out of four TExEval domains. This does not affect negatively the structural properties of these taxonomies, as they also improve in terms of MD and are only slightly behind in AD in some domains. The case of the Science domain is remarkable, where the automatic EXTASEM! taxonomy shows greater AD than the gold stan-

	FOOD		SCIENCE		EQUIPMENT		CHEMICAL		TERRORISM		ARTIFICIAL INTELLIGENCE	
	TEXE	EXTASEM!	TEXE	EXTASEM!	TEXE	EXTASEM!	TexE.	EXTASEM!	Luu Ahn et al	EXTASEM!	Velardi et al	EXTASEM!
NODES	1556	3647	452	2124	612	2062	17584	4932	123	510	2388	1556
EDGES	1587	3930	465	2243	615	2214	24817	5355	243	548	2386	1610
C.C	1556	3647	452	2124	612	2062	17584	4932	N.A	510	2386	1556
I.N	69	1980	53	611	57	995	3349	2051	N.A	292	747	730
MD	6	9	5	8	6	9	18	8	N.A	7	13	7
AD	3.8	3.6	3.7	3.9	3.6	3.5	9	3.7	N.A	3.4	6.7	3.5

Table 3: Taxonomy structure results.

dard. The one domain that poses most difficulties for our approach is Chemical due to the low coverage this domain has in BABELNET.

As for comparison against automatic taxonomies, while AD and MD are lower than Velardi et al.'s *OntoLearn Reloaded*, note that in their approach many upper-level (not domain-specific) nodes are introduced, which are described as "general enough to fit most domains"[8]. Finally, our evaluation suggests that the Terrorism taxonomy in (Luu Anh, Kim, and Ng 2014) does not have all the components connected. We therefore report statistics on its biggest connected subgraph. Additionally, since it was not constructed on the back of an umbrella root node, we do not report numbers on depth. This reflects the complexity of the taxonomy learning task, where perfectly valid domain-specific taxonomies may be shaped as trees or as directed acyclic graphs, with or without root nodes. Full domain-wise details are provided in Table 3.

Experiment 2 - Hypernym Extraction

We considered WIBI as our main competitor in the task of hypernym extraction due to the similarities in terms of (hyponym, hypernym) extraction from a definition setting.

For each domain, two experts were presented with 100 randomly sampled terms and two possible hypernyms, the hypernym selected by EXTASEM! and the one from WIBI. Each pair was shuffled to prevent evaluators from guessing which could be the source. For each pair of hypernym candidates, evaluators had to decide which of the two options constituted a *valid* hypernym in the given domain. They were allowed to leave this field blank for both systems. If both the hypernyms in WIBI and EXTASEM! were valid, evaluators were asked to decide which system offered the *best* hypernym (or both if it was the same), and for this we asked them to consider the hypernym's semantic relatedness and closeness to the hyponym, as well as relevance to the domain. For example, for the hyponym CHUPA CHUPS, we would prefer LOLLIPOP over COMPANY in the Food domain, even if strictly speaking both options would be valid. We computed inter-rater agreement with the Cohen's Kappa metric over the *valid* and *best* classes, with average results of 0.53 and 0.36.

The results in Table 4 suggest that in general the hypernyms extracted with our procedure are better, i.e. more appropriate to the domain and more informative, than the ones extracted from the syntactically-motivated heuristic described in (Flati et al. 2014).

	FOOD		SCIENCE		EQUIPMENT	
	Valid	Best	Valid	Best	Valid	Best
WIBI	0.85	0.29	0.85	0.39	0.84	0.3
EXTASEM!	**0.94**	**0.91**	**0.91**	**0.83**	**0.90**	**0.83**
	CHEMICAL		AI		TERRORISM	
	Valid	Best	Valid	Best	Valid	Best
WIBI	**0.75**	0.03	0.76	0.39	**0.79**	0.24
EXTASEM!	0.64	**0.32**	**0.84**	**0.80**	0.78	**0.73**

Table 4: Human judgement on the quality of the hypernymic relations provided by WIBI and EXTASEM! for 6 domains.

Conclusion and Future Work

We have presented EXTASEM![9], a system that constructs a domain-specific semantically rich taxonomy from an input terminology. It consists of three main modules, namely: (1) Domain Definition Harvesting, where BABELNET and WIBI are leveraged in order to obtain a significant amount of definitional evidence; (2) Hypernym Extraction and Decomposition, based on a CRF-based sequential classifier and a syntactically-motivated hypernym decomposition algorithm; and (3) Path Disambiguation and Graph Induction, on the back of SENSEMBED, a state-of-the-art vector space representation of individual word senses.

Parting ways from previous approaches in which is-a relation evidence was gathered from non curated data like the web or OIE systems, EXTASEM! explicitly tackles the semantics of each candidate (hyponym, hypernym) pair, as well as its pertinence to the target domain.

EXTASEM! achieves state-of-the-art performance in reconstructing gold standard taxonomies, and is able to extend them retaining their domain relevance. Still, we acknowledge limitations in terms of lackluster coverage for certain domains (e.g. Chemical), as well as potential errors introduced in the hypernym extraction, syntactic decomposition and path weighting modules.

[8]Some of these nodes are *abstraction*, *entity*, *event* or *act*.

[9]This work is partially funded by Dr. Inventor (FP7-ICT-2013.8.1611383) and the SKATER-TALN UPF project (TIN2012-38584-C06-03).

References

Alfarone, D., and Davis, J. 2015. Unsupervised learning of an is-a taxonomy from a limited domain-specific corpus. In *Proceedings of IJCAI 2015*, 1434–1441.

Bohnet, B. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *COLING*, 89–97.

Bordea, G.; Buitelaar, P.; Faralli, S.; and Navigli, R. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation*.

Camacho-Collados, J.; Pilehvar, M. T.; and Navigli, R. 2015. NASARI: A Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, 567–577.

Espinosa-Anke, L.; Saggion, H.; and Ronzano, F. 2015. Weakly supervised definition extraction. In *Proceedings of RANLP 2015*, 176–185.

Etzioni, O.; Cafarella, M.; Downey, D.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D. S.; and Yates, A. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence* 165(1):91–134.

Flati, T.; Vannella, D.; Pasini, T.; and Navigli, R. 2014. Two is bigger (and better) than one: the wikipedia bitaxonomy project. In *ACL*.

Fountain, T., and Lapata, M. 2012. Taxonomy induction using hierarchical random graphs. In *Proceedings of NAACL*, 466–476. Association for Computational Linguistics.

Fowlkes, E. B., and Mallows, C. L. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* 78(383):553–569.

Francom, J., and Hulden, M. 2008. Parallel multi-theory annotations of syntactic structure. In *Proceedings of LREC*.

Glickman, O.; Dagan, I.; and Koppel, M. 2005. A probabilistic classification approach for lexical textual entailment. In *Proceedings of the National Conference On Artificial Intelligence*, volume 20, 1050.

Harabagiu, S. M.; Maiorano, S. J.; and Pasca, M. A. 2003. Open-domain textual question answering techniques. *Natural Language Engineering* 9(03):231–267.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, 539–545. Association for Computational Linguistics.

Hjelm, H., and Buitelaar, P. 2008. Multilingual evidence improves clustering-based taxonomy extraction. In *ECAI*, 288–292.

Hwang, S. J.; Grauman, K.; and Sha, F. 2012. Semantic kernel forests from multiple taxonomies. In *Advances in Neural Information Processing Systems*, 1718–1726.

Iacobacci, I.; Pilehvar, M. T.; and Navigli, R. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL*. Beijing, China: Association for Computational Linguistics.

Kozareva, Z., and Hovy, E. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of EMNLP*, 1110–1118. Association for Computational Linguistics.

Kozareva, Z.; Hovy, E. H.; and Riloff, E. 2009. Learning and evaluating the content and structure of a term taxonomy. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, 50–57.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Liu, X.; Song, Y.; Liu, S.; and Wang, H. 2012. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1433–1441. ACM.

Luu Anh, T.; Kim, J.-j.; and Ng, S. K. 2014. Taxonomy construction using syntactic contextual evidence. In *EMNLP*, 810–819.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, 746–751.

Miller, G. A.; Beckwith, R.; Fellbaum, C. D.; Gross, D.; and Miller, K. 1990. WordNet: an online lexical database. *International Journal of Lexicography* 3(4):235–244.

Navigli, R., and Ponzetto, S. P. 2010. BabelNet: Building a very large multilingual semantic network. In *ACL*, 216–225.

Navigli, R., and Velardi, P. 2010. Learning word-class lattices for definition and hypernym extraction. In *ACL*, 1318–1327.

Navigli, R.; Velardi, P.; and Faralli, S. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI*, 1872–1877.

Navigli, R.; Velardi, P.; and Ruiz-Martínez, J. M. 2010. An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of LREC'10*.

Neelakantan, A.; Shankar, J.; Passos, A.; and McCallum, A. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1059–1069. Doha, Qatar: Association for Computational Linguistics.

Ponzetto, S. P., and Strube, M. 2008. Wikitaxonomy: A large scale knowledge resource. In *ECAI*, volume 178, 751–752.

Saggion, H. 2004. Identifying Definitions in Text Collections for Question Answering. In *Proceedings of LREC*.

Snow, R.; Jurafsky, D.; and Ng, A. Y. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006*, 801–808.

Suchanek, F. M.; Kasneci, G.; and Weikum, G. 2007. Yago: A core of semantic knowledge. In *WWW*, 697–706. ACM.

Velardi, P.; Faralli, S.; and Navigli, R. 2013. OntoLearn Reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics* 39(3):665–707.